Predicting Pitcher Injuries

By: Kurt Bullard, Jake Meagher, and Declan Garvey

Introduction:

We set out to create a model—using logistic regression—that could help predict whether or not a pitcher would get injured the following season based on traditional and advanced statistics from both the previous year and the entire career of each pitcher.

This model could assist MLB front offices in targeting or avoiding certain pitchers in free agency and in trades, or at least adjusting their evaluations of them accordingly, which could dramatically improve team performance from season-to-season. Each year, general managers throughout Major League Baseball dish out tens of millions of dollars to acquire star players during the annual free agency period. Usually, the major "prizes" of free agency are star pitchers who can help bolster a pitching rotation. This year, the Boston Red Sox inked David Price to a seven-year, $217 million contract. Days later, on Dec. 4, 2015, the Arizona Diamondbacks signed star pitcher Zack Greinke to a six-year, $206 million contract. Teams spend a lot of money on these hurlers, so it's imperative that they make sure these investments are sound. Evaluating pitchers for injury risk, therefore, is of the utmost interest for each team, as most of the money designated in these contracts is guaranteed, and an injury would spoil the investment. Over the past five years, over 23 percent have been placed on the Disabled List (DL).

Methods:

Data Collection:

We looked at pitcher data from 2010-2014 to match the respective injury data that we found from 2011-2015, since we were concerned with how last year's usage and performance might affect next year's injury risk. We wanted to look at pitcher-specific injuries—ones that came as a result of pitching stress put on certain parts of the body, as opposed to injuries that were "fluky." For that reason, we only considered injuries involving the arm, shoulder, back, and side. Other injuries, we assumed, were not inherently related to pitching stress (e.g. gastrointestinal). In total, there were 3330 pitcher-seasons that met this initial criteria.

Our independent variables came from three different sources:

Baseball Reference: Strikeouts, Age, Dummy Variable for AL/NL, Games, Games Started, Dummy Variable for Starting Pitcher (Games Started > 0), Complete Games, Innings Pitched, Hits, Runs, BB, FIP, Batters Faced, Strike Percentage, and Career Batters Faced

Baseball Info Solutions Data (from Fangraphs): Percentage of Pitch Thrown and Average Velocity for the Following Pitches: Fastball, Cutter, Slider, and Curveball

Tommy John Database: A dummy variable signaling whether or not a pitcher had had Tommy John Surgery before

We did end up trimming some of the data that we did not find representative of a decent MLB pitcher. For one, we did not include three pitcher seasons where the pitcher made an appearance did not record an out, which messes with FIP and makes the seasons unusable. Also, if a pitcher doesn't record an out the entire year, he's more likely than not a mediocre talent that should not influence the model.

In addition, we did not include pitchers who recorded less than 10 innings in a season. We realize that this may be a bit problematic in that some pitchers may have gotten injured less than 10 innings into the season and did not play for that reason rather than not play due to lack of talent—seven percent of these pitchers ended up getting injured. That being said, the overall injury rate for pitchers hovered around 23%, so most of the pitchers who pitched fewer than 10 innings were simply under-utilized rather than injured. However, there was very little Baseball Info Solutions data for pitchers with few appearances, so it would have thrown off the actual impact of pitch speed and selection. At the end, we were left with 2749 pitcher-seasons.

Variable Selection:

Among the variables we elected to eliminate from the regression model were innings pitched, which showed signs of multicollinearity with batters faced (correlation greater than **.99**). We did keep batters faced in the model though. We also did not include games started, as it showed strong multicollinearity with batters faced (**0.941).**

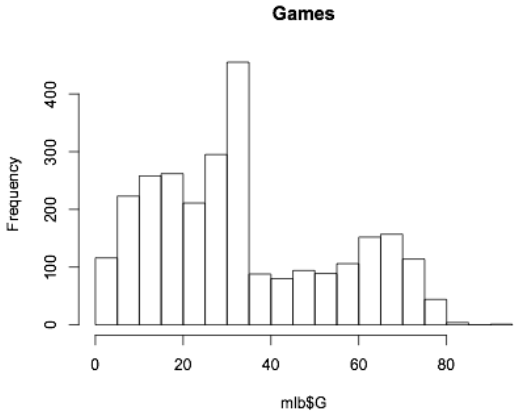
Another variable we did away with was hit by pitches, as HBP and wild pitches had also shown signs of multicollinearity (**.43**). We kept wild pitches, but we did transform it (see below). Hits and runs showed collinearity (**.97**), so we decided to keep hits, because we hypothesize those were more stable across years than were runs, which is based more on the clustering of hits.

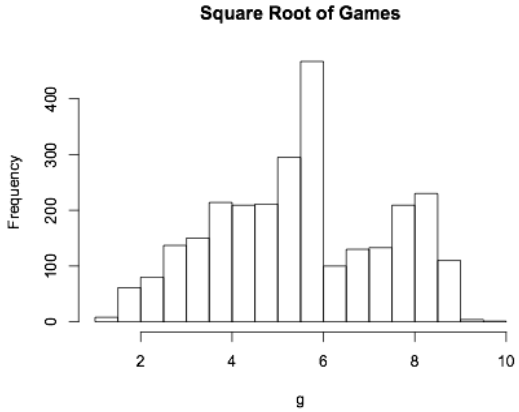
We also noticed collinearity between three variables—strikeouts, hits, and walks—and batters faced (**.93, .98, .89**), which makes sense, since the more you play, the more hits and walks you let up and strikeouts you can register, regardless of skill level. Thus, we transformed the three counting statistics (SO, H, BB) into rates, comparing them each to batters faced (SO/BF, H/BF, BB/BF). As a result, we were able to eliminate the strong correlation between the aforementioned variables and batters faced with these new statistics. We also translated wild pitches to a rate parameter for similar reasons, even though the collinearity between WP and BF was not strong.

We also noticed a relationship between FIP and the rate parameters of SO, H, and BB, so we deleted FIP from the model (**-0.5897501, 0.4537724, 0.321768**). This correlation makes sense, because FIP is a statistic created strictly with SO, BB, and HR.

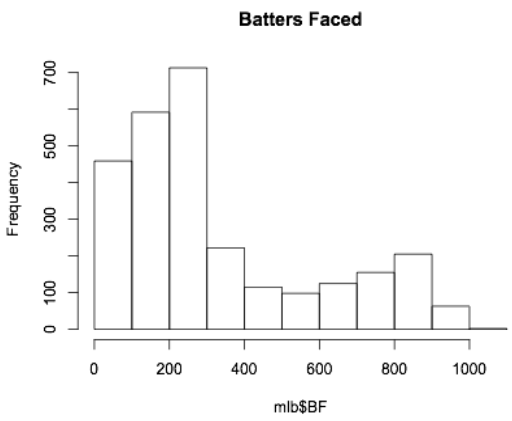
Transformations:

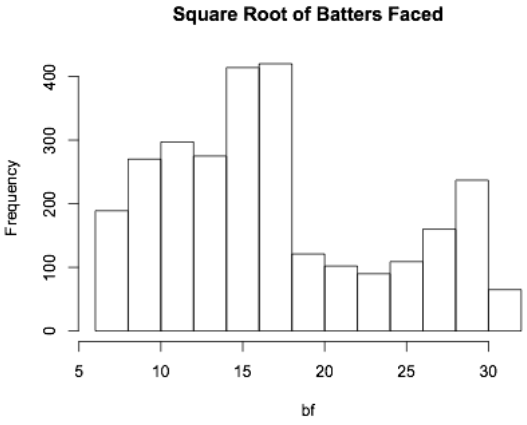
In addition to removing some superfluous variables from the mix, we also decided to transform a handful of others after examining their distributions.



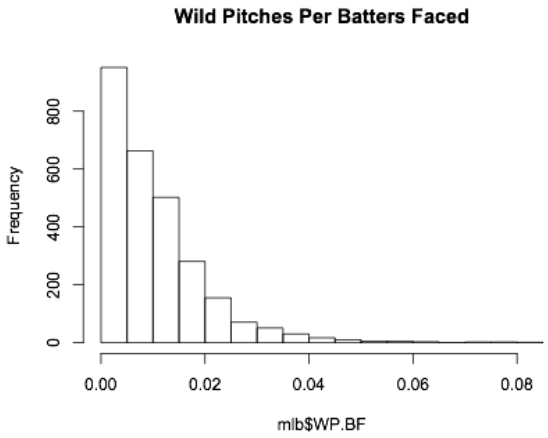


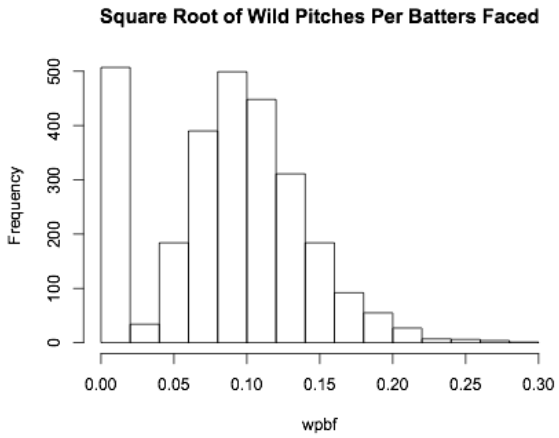
Since the distribution of games was skewed right (see left), we used a square root transformation to increase its normality (see right).



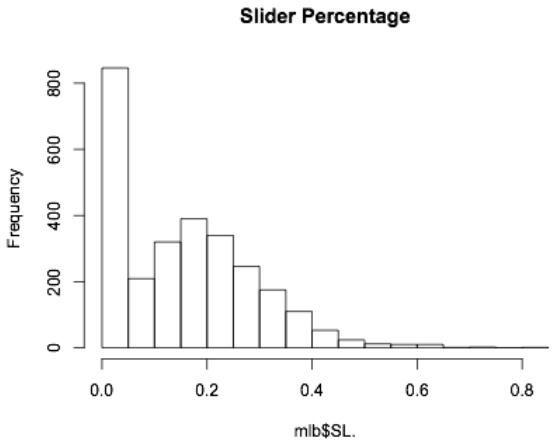


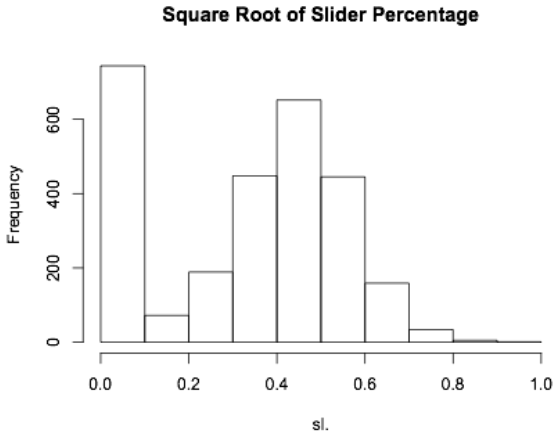
We also transformed batters faced, which also had a right-skewed distribution. Consequently, we utilized another square root transformation to increase its symmetry and normality.



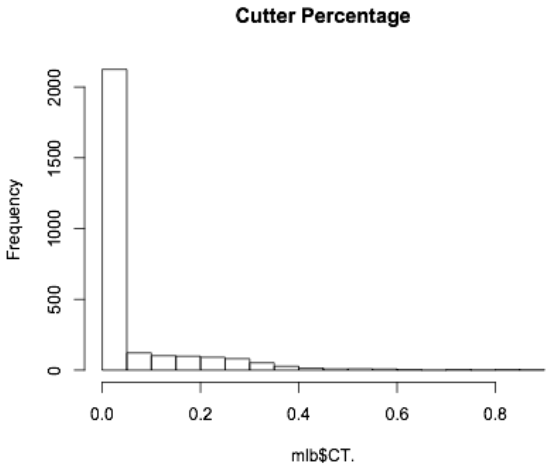


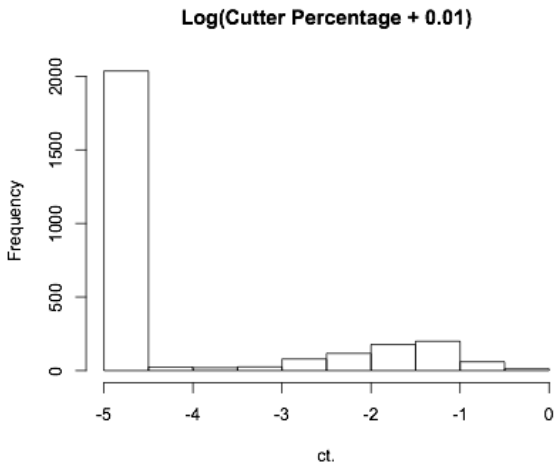
We also transformed wild pitches per batters faced, which had a right-skewed distribution. Thus, we utilized another square root transformation to increase its symmetry and normality. The newly transformed distribution still shows a potential problem as there seems to be a lot of remaining zero-values (or near-zero values), but we dealt with that later on, as we did for the subsequent transformations as well.



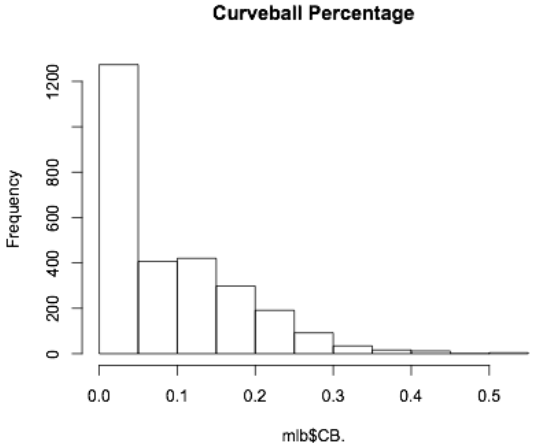


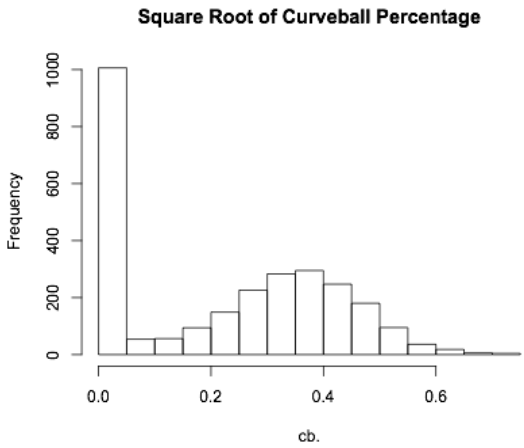
For slider percentage, we see another right-skewed distribution and use a square root transformation to increase its symmetry and normality. We created an indicator variable to interact with the percentage and velocity data, as not all pitchers have every pitch type in their arsenal. This should nullify the effect of the zero-values.



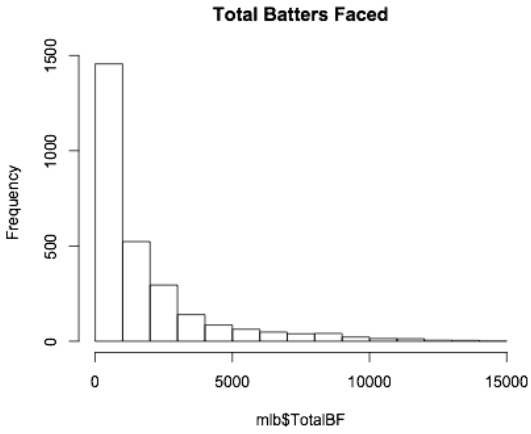


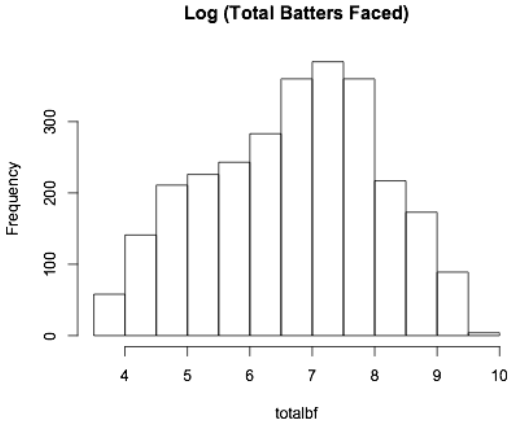
For cutter percentage, we see a very right-skewed distribution and elect to use a log transformation. More specifically, we actually log “Cutter Percentage + 0.01” to account for zero values, because you cannot log zeroes. We created an indicator variable to interact with the percentage and velocity data, as not all pitchers have every pitch type in their arsenal. This should nullify the effect of the zero-values.





For curveball percentage, we see a right-skewed distribution and use a square root transformation to increase its normality. We created an indicator variable to interact with the percentage and velocity data, as not all pitchers have every pitch type in their arsenal. This should nullify the effect of the zero-values.





For total batters faced, we see a very right-skewed distribution and elect to use a log transformation to improve symmetry and normality.

The distributions for the remainder of the variables appeared to be approximately normal, so we found no need to transform them. Below is a key of the variables that we were considering, post-transformations

:**Results:**

Model 1:

In total, out of the 2749 pitcher-seasons in the sample, 644 pitchers were injured the following season, a percentage of 23.4%. Using that data and the variables above, we ran a stepwise logistic regression that yielded the following results:

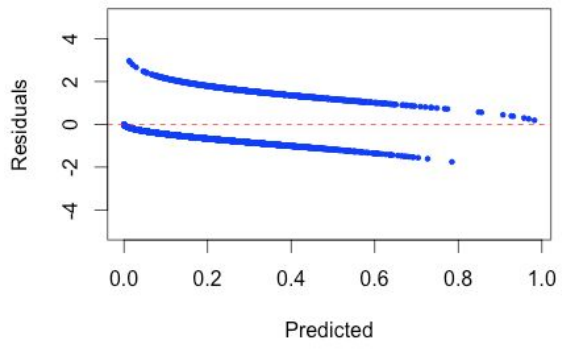
```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.75283  -0.76194  -0.52312  -0.09566   2.96523

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.725e+01  4.091e+01  -0.666 0.505337
sobf        -7.886e+00  2.914e+01  -0.271 0.786650
hbf         -1.354e+01  9.097e+00  -1.489 0.136543
bbbf        -1.138e+01  3.372e+00  -3.375 0.000739 ***
wpbf        -1.036e+01  5.826e+00  -1.779 0.075293 .
age         -1.131e+00  3.106e-01  -3.643 0.000270 ***
sp          -3.523e+00  9.699e-01  -3.633 0.000281 ***
g            1.045e+00  7.690e-01   1.359 0.174212
cg           2.243e+00  6.600e-01   3.399 0.000676 ***
bf           1.024e-01  9.638e-02   1.062 0.288250
strpct       1.177e+02  5.840e+01   2.016 0.043787 *
tj          -2.667e-02  1.639e-01  -0.163 0.870721
fb.          1.306e+00  3.938e+00   0.332 0.740164
fbv          9.469e-01  3.720e-01   2.546 0.010912 *
ct.          8.201e+00  4.006e+00   2.047 0.040621 *
ctv         -5.965e-03  1.802e-02  -0.331 0.740618
sl.          5.629e+00  1.533e+00   3.671 0.000241 ***
slv         -7.744e-03  6.933e-03  -1.117 0.263966
cb.          2.248e+01  1.641e+01   1.370 0.170747
cbv         -3.008e-02  2.281e-02  -1.319 0.187321
chv         -1.991e-01  5.837e-02  -3.412 0.000645 ***
ch.         -3.030e+00  5.157e+00  -0.588 0.556797
totalbf      7.101e-01  1.300e-01   5.463 4.68e-08 ***
chi         -4.650e+00  2.561e+00  -1.816 0.069428 .
sli         -7.511e+00  2.300e+00  -3.266 0.001089 **
cti         -3.370e+01  1.700e+01  -1.982 0.047477 *
age:g        1.604e-02  8.888e-03   1.805 0.071069 .
age:fb.      1.849e-01  9.955e-02   1.858 0.063214 .
sp:g        -5.117e-01  1.581e-01  -3.236 0.001211 **
sp:bf        2.077e-01  6.945e-02   2.990 0.002786 **
tj:ctv      -1.557e-01  1.015e-01  -1.534 0.124908
ct.:cti     -7.311e+00  4.002e+00  -1.827 0.067733 .
slv:sli      5.670e-02  2.637e-02   2.150 0.031549 *
cb.:cbi     -2.936e+01  1.514e+01  -1.939 0.052511 .
cbv:cbi      2.605e-02  1.468e-02   1.775 0.075943 .
sobf:hbf     9.359e+01  2.218e+01   4.219 2.46e-05 ***
cg:bf       -5.294e-02  2.084e-02  -2.540 0.011070 *
fbv:chv      8.078e-04  4.946e-04   1.633 0.102408
sobf:bf     -4.523e-01  1.663e-01  -2.720 0.006524 **
strpct:fbv  -1.623e+00  5.983e-01  -2.713 0.006660 **
ctv:cb.     -6.092e-02  1.418e-02  -4.296 1.74e-05 ***
sobf:fbv     3.742e-01  1.873e-01   1.998 0.045729 *
bbbf:sp      1.654e+01  4.292e+00   3.853 0.000117 ***
strpct:chv   1.760e-01  5.770e-02   3.051 0.002282 **
wpbf:chv     7.716e-02  3.095e-02   2.493 0.012657 *
ct.:ch.     -1.320e+00  6.515e-01  -2.026 0.042754 *
sp:ctv       6.143e-03  3.275e-03   1.876 0.060708 .
tj:cti       1.382e+01  8.855e+00   1.560 0.118647
age:strpct   1.345e+00  4.738e-01   2.839 0.004522 **
strpct:cb.   2.605e+01  1.009e+01   2.583 0.009794 **
sobf:cb.    -2.921e+01  1.003e+01  -2.914 0.003572 **
hbf:cbv     -1.023e-01  4.596e-02  -2.227 0.025970 *
sobf:strpct -5.545e+01  3.261e+01  -1.701 0.088990 .
bf:ch.       4.097e-01  1.347e-01   3.041 0.002354 **
cbv:cti      1.170e-02  6.366e-03   1.838 0.066063 .
ct.:sli     -8.095e-01  2.153e-01  -3.760 0.000170 ***
fb.:sl.     -6.182e+00  2.245e+00  -2.754 0.005888 **
slv:cti      1.427e-02  7.488e-03   1.906 0.056667 .
wpbf:cg     -4.020e+00  1.909e+00  -2.106 0.035214 *
cg:chv      -4.955e-03  2.616e-03  -1.894 0.058231 .
```

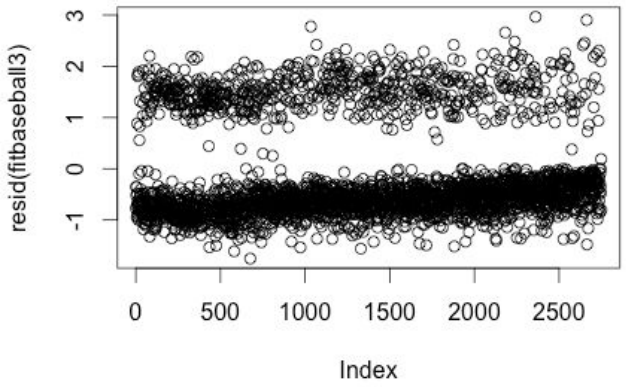
```
bf:totalbf  -3.359e-02  8.490e-03  -3.956 7.61e-05 ***
hbf:wpbf     3.945e+01  2.424e+01   1.627 0.103668
ch.:totalbf -1.272e+00  6.971e-01  -1.825 0.068018 .
chv:chi      4.586e-02  2.939e-02   1.560 0.118744
g:strpct    -1.892e+00  1.166e+00  -1.623 0.104671
hbf:fb.     -1.853e+01  1.155e+01  -1.604 0.108652
fb.:chi      1.926e+00  1.020e+00   1.887 0.059101 .
sp:totalbf   3.546e-01  1.359e-01   2.610 0.009061 **
sobf:cbv     7.673e-02  5.235e-02   1.466 0.142738
wpbf:sl.    -6.418e+00  4.440e+00  -1.445 0.148320
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2993.0  on 2748  degrees of freedom
Residual deviance: 2598.2  on 2679  degrees of freedom
AIC: 2738.2

Number of Fisher Scoring iterations: 7
```

There are a lot of variables included in this model. Some of them make a lot of sense: complete games and career batters faced are both very significant and positive, since they both induce more stress on the arm. However, this model includes several meaningless, yet significant, interactions, like hits per batters faced and curveball velocity, or wild pitches per batters faced and changeup velocity. These interactions led us to question the model's legitimacy, despite the many variables that did make sense.

The Akaike information criterion (AIC) of this model, which measures its quality, is 2738.2, which will be relevant later when we test another model.

When visualizing the residual plot of our model, the resulting graph looked like this:





Although at first we were concerned by the discrete linearity of the plots, we quickly realized that the above graphs was not necessarily an indicator of a bad model. Because, in a logistic regression, the outcome is categorical (can only take on 0 or 1), the residuals for a non-injured pitcher can only be negative, and the residual for an injured pitcher can only be positive. With respect to the first plot, because predicted values and residuals must sum to either zero or one for each observation, the residual plot therefore follows a linear pattern. Nonetheless, because of the high amount of meaningless variables mentioned above, we looked to create a new model.

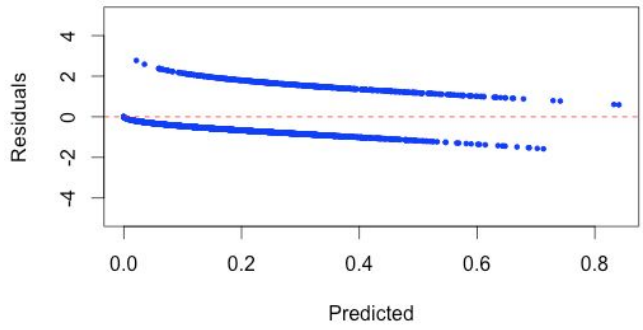
Model 2:

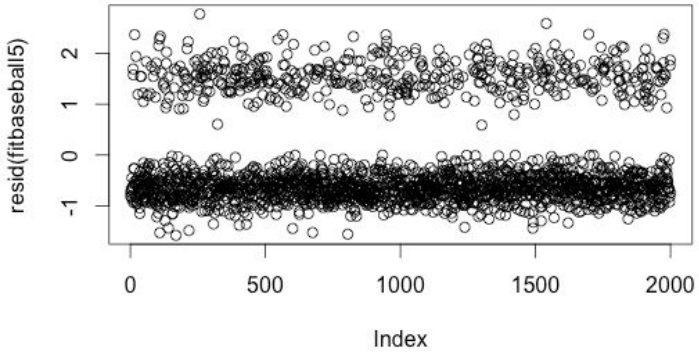
One concern that arose with our first model was the effect of the multitude of zeroes in the cutter percentage, slider percentage, changeup percentage, and curveball percentage predictor variables (see Transformations section). We harbored fear that the skewness of the data was throwing off the accuracy of our model. So, in this model, we only used fastball percentage and fastball velocity, and substituted indicator variables, rather than percentages and velocities, for each of the four offspeed pitches. Much of the same information is still included, as the complement of fastball percentage is offspeed percentage, and fastball velocity is also fairly correlated with the velocity of a pitcher's other pitches. We set the threshold for "having a pitch" at 1 percent, as some pitchers occasionally throw pitches that they're not accustomed to throwing regularly.

The stepwise regression produced the following model:

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -51.818968  37.874618  -1.368 0.171258
sobf        -52.175493  18.561833  -2.811 0.004940 **
hbf         -22.276874   5.099721  -4.368 1.25e-05 ***
bbbf         -9.026819   3.365303  -2.682 0.007311 **
wpbf         -5.401862   2.153145  -2.509 0.012113 *
age          -0.494707   0.298055  -1.660 0.096958 .
sp           -4.105002   3.021735  -1.358 0.174308
g             0.584255   0.226317   2.582 0.009835 **
cg            2.183988   0.611550   3.571 0.000355 ***
bf            0.047967   0.089859   0.534 0.593479
strpct       88.080978  60.587555   1.454 0.146007
fb.           1.835516   0.883123   2.078 0.037669 *
fbv           0.864838   0.372324   2.323 0.020189 *
cti           1.187822   0.334722   3.549 0.000387 ***
sli          -7.133939   2.994876  -2.382 0.017217 *
cbi          -5.392614   2.417127  -2.231 0.025681 *
chi          -8.670173   2.702054  -3.209 0.001333 **
totalbf       0.796275   0.148393   5.366 8.05e-08 ***
rbf         -52.591030  33.119457  -1.588 0.112305
sp:g         -0.458989   0.151184  -3.036 0.002398 **
sp:bf         0.221273   0.066089   3.348 0.000814 ***
sobf:hbf     89.027518  21.877292   4.069 4.71e-05 ***
cg:bf        -0.053967   0.019813  -2.724 0.006454 **
strpct:fbv   -1.514604   0.598678  -2.530 0.011409 *
sobf:bf      -0.631305   0.160809  -3.926 8.64e-05 ***
sobf:fbv      0.489266   0.180899   2.705 0.006838 **
g:totalbf    -0.052132   0.028414  -1.835 0.066541 .
cti:cbi      -0.728756   0.271680  -2.682 0.007310 **
bbbf:sp      15.015629   4.108251   3.655 0.000257 ***
strpct:chi   12.755621   4.186212   3.047 0.002311 **
wpbf:chi      6.158703   2.252664   2.734 0.006258 **
cti:sli      -0.632725   0.289052  -2.189 0.028599 *
strpct:rbf  129.797458  52.511735   2.472 0.013444 *
age:rbf      -1.087107   0.413394  -2.630 0.008546 **
sp:rbf       -8.033157   3.568559  -2.251 0.024380 *
wpbf:cg      -4.587390   1.868486  -2.455 0.014083 *
strpct:cbi    8.351543   3.750254   2.227 0.025952 *
bf:totalbf   -0.016574   0.007774  -2.132 0.033007 *
age:strpct    0.865947   0.446619   1.939 0.052514 .
cg:chi       -0.310582   0.162082  -1.916 0.055339 .
wpbf:cbi      3.776328   2.080617   1.815 0.069523 .
fb.:sli      -2.426746   1.049715  -2.312 0.020788 *
fbv:sli       0.096830   0.034553   2.802 0.005073 **
sp:fbv        0.038132   0.032722   1.165 0.243890
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2993.0  on 2748  degrees of freedom
Residual deviance: 2682.2  on 2705  degrees of freedom
AIC: 2770.2

Number of Fisher Scoring iterations: 7
```

The AIC in this model is **2770.2,** which is slightly higher than the first model that we tested, *signaling that the first model is perhaps better.* The lowest AIC doesn't guarantee the best model, but it often leads to a more useful model.





The residual plot follows a linear pattern—as was explained above, that is expected in a logistic regression based on the binary outcome.

Cross Validation:

After running 2,000 simulations of cross-validation, training the data using 2,000 observations to predict the other 749, we found that the average Sum of Squared Error for each model was as follows:

Model 1	Model 2
4.986778	4.676602

The lower SSE average in Model 2 suggests, contrary to the AIC, that Model 2 may have more predictive power.

Since the two measures of model quality were contradicting, we had to decide which model to use going forward (inevitably, neither became our final model). We chose the first model as a result of its lower AIC, even though it produced a higher SSE in the cross-validation process. In addition, a lot of the pitch selection-related variables were significant (slider

percentage, cutter velocity, etc.). So, we felt like it was critical to include a model that contained these variables.

Predictions for 2016 Season (Attempt 1):

Last year, 598 pitchers threw 10 or more innings. We had to remove 19 pitchers because they did not have Baseball Info Solutions data, which left 579 pitchers from last season who threw 10 or more innings and had the relevant velocity and pitch selection data.

When we ran our predictions, however, we noticed that a lot of pitchers had either a near-zero or near-1 probability of getting injured next season, which does not make sense. We then noticed, after looking through the predictions more closely, that relief pitchers exhibited a 100% probability of getting injured, while starting pitchers had a near-zero percent chance of getting injured. On their own, these results do not make sense. But particularly since starters throw more than relievers do, these predictions are especially puzzling. **For this reason, we cannot go forward with this model. We have to make a new model that gives better predictions.**

Model 3:

We believe that the aforementioned problem arose due to overfitting —the model was not generic enough to apply to a new dataset. In this third model, we eliminate some variables that we do not believe are critical to the regression for fear of overfitting. We eliminated games, league, wild pitches per batters faced, hits per batters faced, strike percentage, starting pitcher, as well as all of the pitch selection data besides fastball percentage, velocity and an indicative variable for cutters. We felt comfortable with the elimination of these aforementioned variables for the following reason:

Games: Although not correlated with batters faced, it is a measure of usage. Batters faced is more specific, so we'll eliminate games.

League: Even though pitchers have to bat in the National League, this probably does not affect pitcher injury (at least the types of injuries we're looking at).

Wild Pitches per BF: We already have a measure of accuracy in walks, so there's not a great need for another variable that represents a similar measure.

Hits per BF: We think that this may not be relevant, since hits per batters faced may already be reflected by BF.

Strike Percentage: We already have a measure of accuracy in walks, so there's not a great need for another variable that represents a similar measure.

SP: In the first model, we found that the starting pitcher indicator variable led to a lot of 0% and 100% values, so we will try to leave out whether or not a pitcher started games or came in later.

Pitch Selection: Fastball percentage also holds in its value the complement, which represents the amount of off-speed thrown, which could be predictive of injury based off arm motion. Cutters have also been hypothesized to lead to injury, so we kept an indicator for that. Lastly, we only use fastball velocity, as we believe that it should be correlated with the velocity of other pitches.

So, we keep only the cutter indicator, fastball velocity and fastball percentage from the pitch selection data.

We also included a few interaction variables: Tommy John and Fastball velocity, complete games and batters faced, fastball percentage and velocity, strikeouts per batters faced and fastball velocity, strikeouts per batters faced and age, age and velocity, walks and tommy john, and finally the cutter indicator and fastball percentage.

After running the stepwise regression, we found the following results:

```
Call:
glm(formula = mlb$Outcome ~ (sobf + bbbf + age + cg + bf + tj +
    fb. + fbv + cti + totalbf + tj:fbv + cg:bf + sobf:bf + fb.:fbv +
    sobf:age + sobf:fbv + age:fb. + bbbf:tj + fb.:cti), family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5946  -0.7701  -0.6243  -0.3054   2.5803

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.268818   2.728079   1.198 0.230834
sobf        -19.025987  13.156689  -1.446 0.148147
bbbf         -1.955044   1.836845  -1.064 0.287171
age          -0.258119   0.067523  -3.823 0.000132 ***
cg            2.175238   0.536625   4.054 5.04e-05 ***
bf            0.092589   0.028605   3.237 0.001208 **
tj            4.243848   1.805301   2.351 0.018735 *
fb.         -14.363413   4.740579  -3.030 0.002446 **
fbv          -0.008378   0.022067  -0.380 0.704199
cti          -0.554864   0.518948  -1.069 0.284975
totalbf       0.268568   0.061817   4.345 1.40e-05 ***
tj:fbv       -0.039445   0.019448  -2.028 0.042541 *
cg:bf        -0.073133   0.018547  -3.943 8.04e-05 ***
sobf:bf      -0.452888   0.140417  -3.225 0.001258 **
fb.:fbv       0.108390   0.038887   2.787 0.005315 **
sobf:age      0.507367   0.202724   2.503 0.012323 *
sobf:fbv      0.163297   0.119396   1.368 0.171407
age:fb.       0.141695   0.093995   1.507 0.131688
bbbf:tj      -6.828134   4.401642  -1.551 0.120837
fb.:cti       1.320874   0.908093   1.455 0.145791
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2993.0  on 2748  degrees of freedom
Residual deviance: 2827.2  on 2729  degrees of freedom
AIC: 2867.2

Number of Fisher Scoring iterations: 5
```

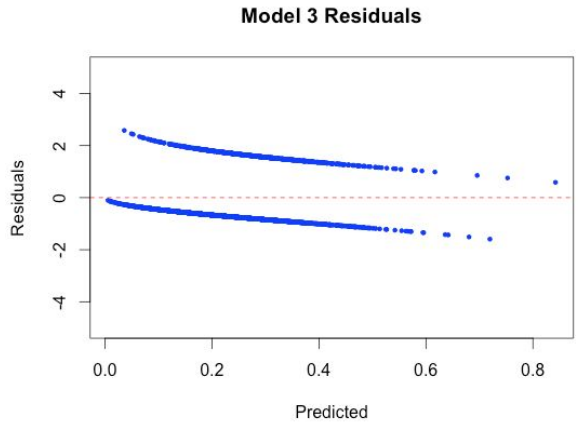
A lot of variables in this model make sense. Throwing more complete games leads to a higher injury rate, as does facing more batters. Having received Tommy John surgery before also points to higher rates of injury, as it makes sense that pitchers who have had arm trouble in the past may have it in the future. Fastball percentage is negatively correlated, which makes sense because the less fastballs you throw, the more offspeed you throw, which means higher stress on the arm. Career total batters faced is also positive, which makes sense because the more you throw, the more of a physical toll it takes on your arm.

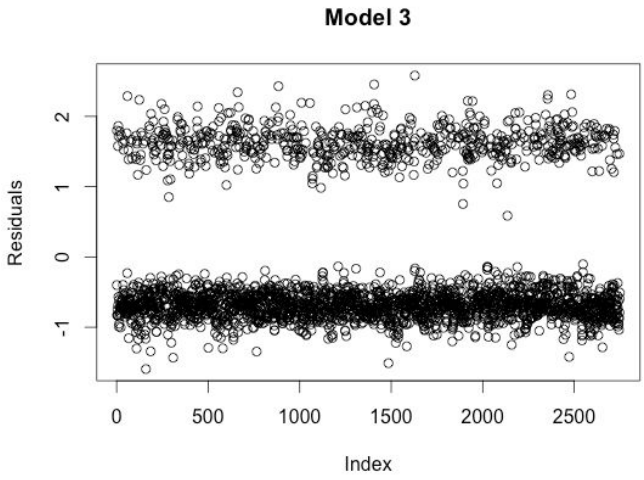
There are also some interaction terms that are significant. Complete games are less impactful when you’ve thrown to more batters, but this may arise because the best pitchers are ones facing the most batters and throwing the most complete games. Strikeouts are also more costly when you get older, as strikeout pitchers tend to throw more pitches in an at-bat than do groundball and flyball pitchers.

One variable in particular stands out as odd: age. It is slightly negative, but this is most likely because of sample bias — the best pitchers who are old probably lasted this long because they're durable, so they'll make it seem like the older you get, the more durable you become.

The AIC of this model is 2867.2, which will be relevant when comparing it to the next model.

Residuals:





The residual plot therefore follows a linear pattern—as was explained above, that is expected in a logistic regression based on the binary outcome.

Model 4:

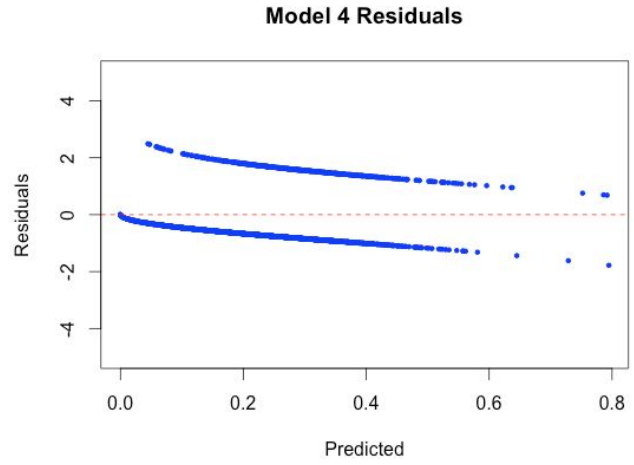
For this model, we added back a few variables just to make sure we didn't overreact to our fear of overfitting. We added back wild pitches per batters faced, hits per batters faced, and strike percentage, and the results were as follows:

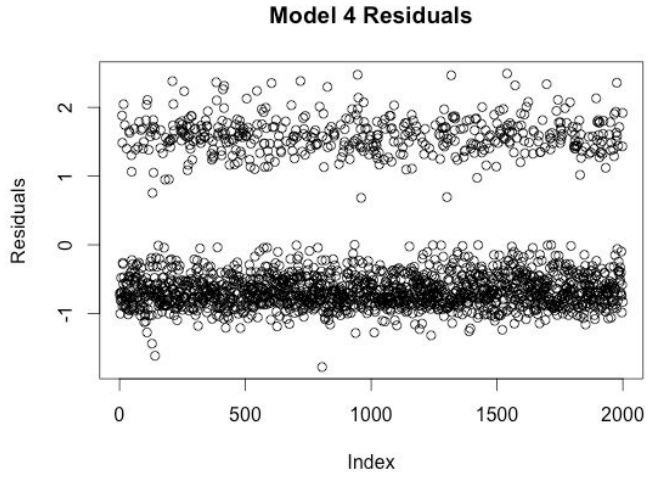
```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6987  -0.7907  -0.6016  -0.1686   2.5569

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -33.925189  32.776299  -1.035 0.300644
sobf        -42.526334  29.356767  -1.449 0.147448
hbf         -28.404603   4.707471  -6.034 1.60e-09 ***
strpct       76.926886  53.656863   1.434 0.151663
bbbf         -6.142614   2.780612  -2.209 0.027169 *
age          -0.862567   0.274084  -3.147 0.001649 **
cg            1.963441   0.549272   3.575 0.000351 ***
bf            0.174610   0.056953   3.066 0.002171 **
tj            5.234814   3.238213   1.617 0.105970
fb.         -12.327106   4.903298  -2.514 0.011936 *
fbv           0.578145   0.339538   1.703 0.088617 .
cti          -0.643722   0.522478  -1.232 0.217928
totalbf       0.476785   0.106882   4.461 8.16e-06 ***
tj:fbv       -0.056465   0.035270  -1.601 0.109391
cg:bf        -0.065719   0.019062  -3.448 0.000566 ***
sobf:bf      -0.528980   0.155068  -3.411 0.000647 ***
fb.:fbv       0.081865   0.041832   1.957 0.050347 .
sobf:age      0.342873   0.233754   1.467 0.142427
sobf:fbv      0.540704   0.197447   2.738 0.006172 **
age:fb.       0.158080   0.090335   1.750 0.080132 .
fb.:cti       1.522808   0.914542   1.665 0.095892 .
sobf:hbf    107.142518  21.330244   5.023 5.09e-07 ***
strpct:fbv   -1.037282   0.552233  -1.878 0.060335 .
strpct:age    0.970587   0.427041   2.273 0.023037 *
sobf:strpct -47.074104  29.430806  -1.599 0.109713
bf:totalbf   -0.010521   0.006241  -1.686 0.091835 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2993  on 2748  degrees of freedom
Residual deviance: 2765  on 2723  degrees of freedom
AIC: 2817

Number of Fisher Scoring iterations: 7
```

The model is similar to Model 3, but there are some interesting variables selected in this model. For one, hits per batters faced is hugely negative, which doesn't make sense, since giving up hits leads to more batters faced. Strikeouts per batters faced with hits per batters faced is also hugely positive and significant, which doesn't make sense for a relatively meaningless interaction. Age and strike percentage is also significant here, despite it being a meaningless interaction.

Even though the AIC is 2817—50 lower than Model 3—it looks like the other model makes more sense and could be a better predictive measure.

Residuals:





The residual plot follows a linear pattern—as was explained above, that is expected in a logistic regression based on the binary outcome.

Cross Validation:

The cross-validation process is already described above, so there is no need to explain it again. The results are as follows:

Model 1:4.98
Model 2: 4.67
Model 3: 3.22
Model 4: 4.23

We see that Model 3 is overwhelmingly better at making predictions than any of the other three models, meaning that this model in most likelihood has the most external validity, even though it does not have the lowest AIC. So, **we will choose Model 3 as our final model to make predictions for next season.**

As a reminder, Model 3 is below:

```
Call:
glm(formula = mlb$Outcome ~ (sobf + bbbf + age + cg + bf + tj +
    fb. + fbv + cti + totalbf + tj:fbv + cg:bf + sobf:bf + fb.:fbv +
    sobf:age + sobf:fbv + age:fb. + bbbf:tj + fb.:cti), family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5946  -0.7701  -0.6243  -0.3054   2.5803

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.268818   2.728079   1.198 0.230834
sobf        -19.025987  13.156689  -1.446 0.148147
bbbf         -1.955044   1.836845  -1.064 0.287171
age          -0.258119   0.067523  -3.823 0.000132 ***
cg            2.175238   0.536625   4.054 5.04e-05 ***
bf            0.092589   0.028605   3.237 0.001208 **
tj            4.243848   1.805301   2.351 0.018735 *
fb.         -14.363413   4.740579  -3.030 0.002446 **
fbv          -0.008378   0.022067  -0.380 0.704199
cti          -0.554864   0.518948  -1.069 0.284975
totalbf       0.268568   0.061817   4.345 1.40e-05 ***
tj:fbv       -0.039445   0.019448  -2.028 0.042541 *
cg:bf        -0.073133   0.018547  -3.943 8.04e-05 ***
sobf:bf      -0.452888   0.140417  -3.225 0.001258 **
fb.:fbv       0.108390   0.038887   2.787 0.005315 **
sobf:age      0.507367   0.202724   2.503 0.012323 *
sobf:fbv      0.163297   0.119396   1.368 0.171407
age:fb.       0.141695   0.093995   1.507 0.131688
bbbf:tj      -6.828134   4.401642  -1.551 0.120837
fb.:cti       1.320874   0.908093   1.455 0.145791
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2993.0  on 2748  degrees of freedom
Residual deviance: 2827.2  on 2729  degrees of freedom
AIC: 2867.2

Number of Fisher Scoring iterations: 5
```

The log-odds of injury risk are negatively correlated with (from most to least significant, where significant relationships are bolded):

- **Complete Games with the Log of Batters Faced**
- **Age**
- **Strikeouts**
- **Fastball Percentage**
- **Tommy John With Fastball Velocity**
- Walks per batters faced and Tommy John
- Strikeouts per batters faced
- Walks per batters faced
- Cutter Indicator

- Fastball Velocity

The log-odds of injury risk are positively correlated with (from most to least significant, where significant relationships are bolded):

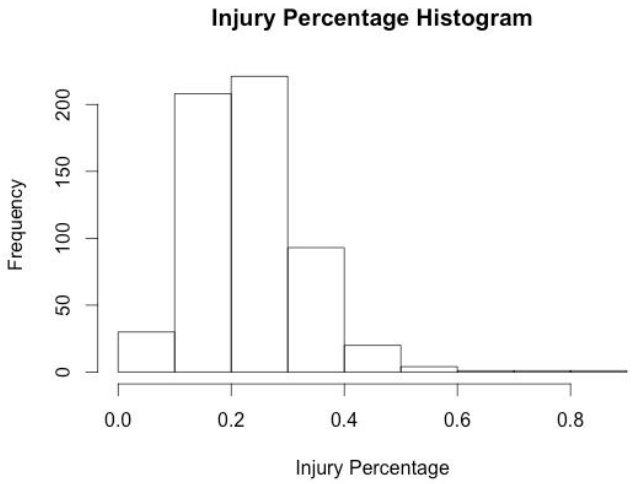
- **Total Batters Faced**
- **Complete Games**
- **Log of Batters Faced**
- **Fastball Velocity with Fastball Percentage**
- **Strikeouts per Batters Faced and Age**
- **Tommy John**
- Fastball Percentage and Age
- Fastball Percentage and Cutter Indicator
- Strikeouts Per Batters Faced and Fastball Velocity

The most significant variables in this model make a lot of sense. In terms of positive correlations, complete games, total batters faced, and the log of the previous season's batters faced all make sense because it punishes you for throwing a lot of pitches in individual games, seasons, and careers. Fastball velocity and percentage also make sense, since a quicker fastball is more dangerous when you throw it more often. Lastly, having had Tommy John points to having had previous injury, which is a good indicator of getting injured once again.

The significant negative correlations also make sense in terms of the structure of the data set. Fastball percentage is negatively correlated with injury risk, as the less fastball one throws, the more arm-straining offspeed pitches are thrown. The best pitchers are the ones who throw the most strikeouts and complete games, meaning they probably have sound mechanics, explaining why strikeouts and complete games with the log of batters faced are significant. Tommy John and Fastball Velocity are significant, which could serve as an indicator to the level of healing from the last injury—perhaps pitchers who recover more fully from it throw faster upon return. Lastly, age is also in the regression, but that is most likely due to sample bias and is a slight shortcoming of our model: the pitchers who lasted to 32-plus were usually the really good ones who in most likelihood never experienced any devastating injury; meanwhile, a lot of really young pitchers get injured and never come back from it.

Predictions for 2016 (Final):

Last year, 598 pitchers threw more than 10 innings. We had to remove 19 pitchers because they did not have Baseball Info Solutions data, which left 579 pitchers from last season who threw 10 or more innings and had the relevant velocity and pitch selection data. The following is the histogram of the injury risk predictions:



We predicted the average risk of a pitcher getting injured in 2016 to be 23.2%, which makes sense, since the five-year mean was in fact 23.4%. We would expect future years to hover around this value, since there hasn't been a secular change in pitcher usage or philosophy.

The range of the model's predictions went from a peak of 80.3% to a low of 2.8%. We present the top 10 most likely and least likely pitchers to get injured next season, respectively.

Top Ten

Name	Injury Risk
Josh Tomlin	80.3%
Rich Hill	78.9%
Carter Capps	61.7%
Nate Jones	56.6%
Brandon McCarthy	53.0%
Adam Ottavino	52.7%
Trevor Rosenthal	51.3%
Aroldis Chapman	48.2%
Wade Davis	47.9%
Andrew Miller	47.3%

Bottom Ten

Name	Injury Risk
Bryan Shaw	2.8%
Scott Atchison	3.9%
Dustin McGowan	5.8%
Eric Stults	5.9%
R.A. Dickey	6.0%
Randy Choate	6.1%
Brad Ziegler	6.1%
Anthony Varvaro	6.2%
Matt Harrison	6.8%
Scott Copeland	6.8%

The pitcher most likely to get injured is Cleveland Starter Josh Tomlin. The starting pitcher is just coming off of Tommy John surgery, and also only throws 53% fastballs, which was in the lowest quartile last season. He's faced 1,675 batters in his career already, and also threw two complete games last year. The pitcher will turn 31 this season as well, which doesn't bode well since he strikes out approximately one every four batters.

High up on the list is Aroldis Chapman, who the Dodgers tried to trade for recently. The Reds' reliever strikes out two out of every five batters, and also throws his fastball an average of 99.5 MPH, which is the highest in the league.

Towards the bottom of the list is R.A. Dickey, which also makes sense, since he is knuckleball pitcher, and that tends to put less stress on the arm than do cutters and curveballs and high-velocity fastballs.

Final Thoughts:

"All models are wrong, but some are useful" - George E. P. Box

We believe that the logistic regression predicting pitcher injuries is a useful model given that it was constructed using only publicly available baseball statistics. However, there is more information that we would want to make a better model that we can't have—eating habits and training regimen, amongst other things that there is no data for. In addition, pitcher mechanics are also a large component of injury risk—those who have worse fundamentals tend to get injured at a higher rate. Using this model in conjunction with qualitative analysis of one's pitching motion would perhaps be even more helpful. Nonetheless, our model is a useful start in identifying pitcher injury risk so that both pitchers themselves and team management can adjust pitching selection and workload as a means of injury prevention.